Stat 8931 (Aster Models)
Lecture Slides Deck 9

Charles J. Geyer

School of Statistics
University of Minnesota

June 7, 2015

LM vs. GLM vs. EFM

GLM and EFM (exponential family models) are mostly like LM.

There are differences.

- In GLM and EFM there is a difference between mean value and canonical parameters. In LM they are the same.
- In GLM and EFM inference is only approximate (large n, asymptotic). In LM inference based on t and F distributions is exact (if you believe the errors are exactly mean zero homoscedastic normal),

But most things are more or less the same.

MLE at Infinity

In this subject, LM and EFM are radically different.

LM can never have MLE "at infinity".

EFM can.

MLE at Infinity (cont.)

Begin with the simplest example.

We observe one Binomial(n, p) random variable x.

MLE for p is $\hat{p} = x/n$.

There are no canonical parameter values corresponding to these "usual" parameter values

$$\theta = \operatorname{logit}(p) = \log(p) - \log(1 - p)$$

does not exist when $p = 0$ or $p = 1$.

$$\begin{aligned} \operatorname{logit}(p) &\to -\infty, && \text{as } p \to 0 \\ \operatorname{logit}(p) &\to +\infty, && \text{as } p \to 1 \end{aligned}$$

we can (loosely speaking) call these MLE "at infinity".

Degeneracy

Binomial(n, p) distributions with $p = 0$ or $p = 1$ are degenerate.

$p = 0$ implies $x = 0$ with probability one.

$p = 1$ implies $x = n$ with probability one.

Exponential families do not have degenerate distributions. Every distribution in the family has the same sets of probability zero, the same support.

So (considered as an exponential family) the binomial family does not contain these degenerate distributions. Hence the MLE does not exist (in the exponential family) when $x = 0$ or $x = n$.

Degeneracy (cont.)

We want to say the MLE is $\hat{p} = 0$ or $\hat{p} = 1$ (respectively) but there is no corresponding $\hat{\theta} = \operatorname{logit}(\hat{p})$.

We could say, let's not use exponential family theory here, but we have to use it for generalized linear models, for log-linear models for categorical data analysis, and for aster models.

This issue has analogs in multiparameter exponential families.

But the high-dimensional geometry is hard to visualize.

Convex Support and Support Function

For any exponential family, the **convex support** of the canonical statistic is the smallest closed convex set that has probability one (all distributions in an exponential family agree on which sets have probability zero or probability one).

Let C be a set in $\mathbb{R}^d$. The **support function** of C is defined by

$$\sigma_C(\delta) = \sup_{y \in C} \langle y, \delta \rangle, \qquad \delta \in \mathbb{R}^d$$

The supremum may be infinite, in which case the value is $+\infty$.

Distributions that are Limits at Infinity

Theorem (Geyer, PhD thesis and *Electronic Journal of Statistics*, 2009). For a full exponential family having dimension d, canonical statistic y, canonical parameter φ, convex support C, canonical parameter space Φ, and PMDF of the canonical statistic f_φ, fix $\delta \in \mathbb{R}^d$, and define

$$H_\delta = \{\, y \in \mathbb{R}^d : \langle y, \delta \rangle = \sigma_C(\delta) \,\}$$

(H_δ is empty if $\sigma_C(\delta) = +\infty$), then for all $\varphi \in \Phi$

$$\lim_{s \to \infty} f_{\varphi + s\delta}(y) = \begin{cases} 0, & \langle y, \delta \rangle < \sigma_C(\delta) \\ f_\varphi(y) / \operatorname{pr}_\varphi(H_\delta), & \langle y, \delta \rangle = \sigma_C(\delta) \\ +\infty, & \langle y, \delta \rangle > \sigma_C(\delta) \end{cases} \qquad (*)$$

where the middle case is interpreted as $+\infty$ if $\operatorname{pr}_\varphi(H_\delta) = 0$.

Distributions that are Limits at Infinity (cont.)

$$\lim_{s \to \infty} f_{\varphi + s\delta}(y) = \begin{cases} 0, & \langle y, \delta \rangle < \sigma_C(\delta) \\ f_\varphi(y) / \operatorname{pr}_\varphi(H_\delta), & \langle y, \delta \rangle = \sigma_C(\delta) \\ +\infty, & \langle y, \delta \rangle > \sigma_C(\delta) \end{cases} \qquad (*)$$

We are only interested in the case $\operatorname{pr}_\varphi(H_\delta) > 0$ when the limit is a PMDF

$$f_\varphi(y \mid H_\delta) = \begin{cases} 0, & \langle y, \delta \rangle < \sigma_C(\delta) \\ f_\varphi(y) / \operatorname{pr}_\varphi(H_\delta), & \langle y, \delta \rangle = \sigma_C(\delta) \\ +\infty, & \langle y, \delta \rangle > \sigma_C(\delta) \end{cases} \qquad (**)$$

The value $+\infty$ in the third case is not a problem because such y are not in the convex support. (This is a convention of measure-theoretic probability: $0 \times \infty = 0$.)

Distributions that are Limits at Infinity (cont.)

Thus we have

$$f_{\varphi + s\delta}(y) \to f_{\varphi}(y \mid H_{\delta}), \qquad \text{as } s \to \infty \text{, for all } y \text{ and } \varphi$$

Pointwise convergence of PMDF implies convergence in distribution but is stronger (actually convergence in total variation).

These conditional distributions, which are also limits of distributions in the original family, are degenerate, concentrated on the hyperplane H_{δ}.

Exponential Family PMDF

The PMDF f_φ can be written

$$f_\varphi(y) = f_{\varphi^*}(y) e^{\langle y, \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)}$$

where c is the cumulant function of the family (deck 2, slides 38–40).

Hence

$$f_\varphi(y \mid H_\delta) = \frac{f_{\varphi^*}(y)}{\text{pr}_\varphi(H_\delta)} e^{\langle y, \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)}$$

Limiting Conditional Model

$$f_{\varphi}(y \mid H_{\delta}) = \frac{f_{\varphi^*}(y)}{\operatorname{pr}_{\varphi}(H_{\delta})} e^{\langle y, \varphi - \varphi^* \rangle - c(\varphi) + c(\varphi^*)}$$

Hence the family of all such limits

$$\mathcal{F}_{\delta} = \{\, f_{\varphi}(\,\cdot \mid H_{\delta}) : \varphi \in \Phi \,\}$$

is another exponential family with canonical statistic y and canonical parameter φ and cumulant function

$$c_{\delta}(\varphi) = c(\varphi) - c(\varphi^*) + \log \operatorname{pr}_{\varphi}(H_{\delta})$$

Conditioning on H_{δ} turns the original exponential family into another exponential family.

Aggregate Exponential Family

In the special case $\delta = 0$ the set H_δ is not a hyperplane but all of $\mathbb{R}^d$ and $\mathcal{F}_\delta$ is just the original family.

The union

$$\bigcup_{\substack{\delta \in \mathbb{R}^d \\ \operatorname{pr}_\varphi(H_\delta) > 0}} \mathcal{F}_\delta \tag{$\star$}$$

in "nice" cases contains the original family and all its limits.

In some "pathological" cases some families $\mathcal{F}_\delta$ are not full and one may need to take limits in them.

In other "pathological" cases the union $(\star)$ does not have all the limits, and one must apply the same limiting procedure to each $\mathcal{F}_\delta$ (and possibly iterate the limiting procedure over and over until all limits are found — since each limiting procedure reduces the dimension of the family by at least one, the recursion stops after at most d steps).

Aggregate Exponential Family (cont.)

It is not obvious that taking limits in straight lines (parameter values $\varphi + s\delta$ and s goes to infinity with φ and δ fixed) gets all possible limits, but Chapter 4 of Geyer (PhD thesis) shows it does (if iterated limits are done).

This process of taking all limits is called the Barndorff-Nielsen completion of the family.

This construction seems complicated (and it is) but it is the price we pay for using exponential family theory.

When MLE do not exist in the original family, they may exist in the Barndorff-Nielsen completion.

Directions of Recession and Constancy

For a regular full exponential family with log likelihood l, canonical statistic Y, and observed value of the canonical statistic y,

we say δ is a **direction of recession** of l if

$$\langle Y, \delta \rangle \le \langle y, \delta \rangle, \qquad \text{almost surely},$$

and we say δ is a **direction of constancy** of l if

$$\langle Y, \delta \rangle = \langle y, \delta \rangle, \qquad \text{almost surely}.$$

Every direction of constancy is a direction of recession.

δ is a direction of constancy if and only if both δ and $-\delta$ are directions of recession.

Directions of Recession and Constancy (cont.)

Consider a regular full exponential family with log likelihood l, observed value of the canonical statistic y, canonical parameter φ, convex support C, and canonical parameter space Φ.

If δ is a direction of recession, then for all $\varphi \in \Phi$

$$\varphi + s\delta \in \Phi, \qquad s \geq 0.$$

If δ is a direction of constancy, then for all $\varphi \in \Phi$

$$s \mapsto l(\varphi + s\delta) \text{ is a constant function on } (-\infty, \infty).$$

If δ is a direction of recession that is not a direction of constancy, then for all $\varphi \in \Phi$

$$s \mapsto l(\varphi + s\delta) \text{ is a strictly increasing function on } [0, \infty).$$

Directions of Recession and Constancy (cont.)

Theorem (Geyer, PhD thesis and 2009). In a full regular exponential family the MLE exists if and only if every direction of recession is a direction of constancy.

This is basically a general fact about concave functions (Rockafellar, *Convex Analysis*, 1970, Theorem 27.1 (b)) applied to exponential families.

Corollary. In a full regular exponential family the MLE exists and is unique if and only if there are no directions of recession (hence no directions of constancy).

One might think we would want uniqueness of MLE guaranteed by corollary, but it turns out that in this context we do not.

Directions of Recession and Constancy (cont.)

A direction δ is a direction of constancy if and only if canonical parameter values $\varphi + s\delta$ correspond to the same probability distribution for all $s \in \mathbb{R}$.

So when there is a direction of constancy δ and $\hat{\varphi}$ is an MLE, then so is $\hat{\varphi} + s\delta$ for all $s \in \mathbb{R}$ but all of these MLE correspond to the same probability distribution.

A direction δ is a direction of constancy (repeating what was said before in different language) if and only if the family is degenerate, concentrated on the hyperplane H_δ.

Before, we ruled out directions of constancy, but now we cannot because all of the distributions added in the Barndorff-Nielsen completion are degenerate, concentrated on some hyperplane H_δ.

Directions of Recession and Constancy (cont.)

Theorem (Geyer, PhD thesis and 2009). If $\hat{\varphi}_1$ and $\hat{\varphi}_2$ are MLE in a regular full exponential family, then $\hat{\varphi}_1 - \hat{\varphi}_2$ is a direction of constancy.

This says that directions of constancy are the only kind of nonuniqueness a regular full exponential family can have.

Hence when the MLE is nonunique, all MLE correspond to the same probability distribution.

Nonuniqueness is not a problem for statistical inference, merely a computational nuisance.

Directions of Recession and Constancy (cont.)

Everything said so far applies to any regular exponential family.

In particular, it applies to unconditional canonical affine submodels of aster models just like it applies to aster models.

The only difference is

- the saturated model has canonical statistic y and canonical parameter φ, whereas
- the submodel has canonical statistic $M^T y$ and canonical parameter β.

Limiting Conditional Model

When we have a direction of recession δ that is not a direction of constancy we have

$$l(\beta) = \log f_\beta(y) < \log f_\beta(y \mid H_\delta) = l_\delta(\beta)$$

Thus, if we maximize the log likelihood l_δ for the limiting conditional model (LCM) we maximize the log likelihood over the model that is the union of the original model and the LCM. If the MLE in the LCM exists, then we are done. That is the MLE in the Barndorff-Nielsen completion.

Directions of Recession and Constancy (cont.)

So how do we find directions of recession and constancy?

Directions of constancy are fairly easy. Mostly they arise from formulas specifying model matrices that are not full rank. The R function `aster` takes care of most cases of that automatically.

Directions of recession that are not directions of constancy are hard. They arise when the observed value of the natural statistic is on the relative boundary of the convex support.

For submodels, the support of $M^T y$ is hard to visualize. Geyer (2009) shows how to use computational geometry software (R package `rcdd`) to find directions of recession. Those methods are slow, and their application to aster models has never been worked out.

Directions of Recession and Constancy (cont.)

The R function `summary.aster` has a dumb methodology for finding directions of recession.

If δ is a direction of recession that is not a direction of constancy, then $l(\varphi + s\delta)$ is a strictly increasing function of s. But this function is bounded above because

$$l(\beta + s\delta) \to \log f(y \mid H_\delta), \qquad \text{as } s \to \infty.$$

Thus both first and second derivatives

$$\begin{aligned} \frac{dl(\beta + s\delta)}{ds} &= (y - \mu(a + M\beta + sM\delta)^T M\delta \\ \frac{d^2 l(\beta + s\delta)}{ds^2} &= -\delta^T M^T I(a + M\beta + sM\delta)^T M\delta \end{aligned}$$

must go to zero as $s \to \infty$.

Directions of Recession and Constancy (cont.)

Thus `summary.aster` looks for null eigenvectors of the Fisher information matrix and reports them as possible directions of recession or constancy.

If `aout` is an object of class `"aster"`, then

```
fred <- eigen(aout$fisher, symmetric = TRUE)
sally <- fred$values < max(fred$values) * info.tol
zapsmall(fred$vectors[, sally])
```

is the code in `aster.summary` that computes these possible directions of recession or constancy.

Directions of Recession and Constancy (cont.)

Because computer arithmetic is inexact (about 16 decimal place precision) one cannot expected computed eigenvalues to be exactly zero. Hence we use a tolerance `info.tol`.

This “test” for directions of recession leads to many false positives.

But it also has revealed many true positives: actual directions of recession that were not directions of constancy. These had to be dealt with. They could not be ignored.

Computer Arithmetic

Computer arithmetic is not exact.

```
> .Machine$double.eps
```

```
[1] 2.220446e-16
```

is the precision or **machine epsilon**, the smallest number that when added to one is greater than one

```
> 1 + .Machine$double.eps / 2 - 1
```

```
[1] 0
```

It makes no sense to test the computer's so-called real numbers for equality (to zero or to anything else).

Computer Arithmetic (cont.)

One can change `info.tol` from its default value

```
> sqrt(.Machine$double.eps)
```

```
[1] 1.490116e-08
```

to something smaller. `1e-9` and `1e-10` are fairly safe. `1e-11` and `1e-12` are getting iffy. Much below that is too close to the machine epsilon.

An error of 1 machine epsilon in one calculation can build up to millions or billions of machine epsilons after millions or billions of operations.

Directions of Recession and Constancy (cont.)

Futzing with `info.tol` gives one (uncertain) way to tell whether putative directions of recession `summary.aster` warns about are real ones. If the warning goes away when `info.tol` is lowered *a little bit*, then there is *probably* (cannot be certain) not a problem.

Looking at the putative direction of recession itself is another (even less certain) way to tell whether putative directions of recession `summary.aster` warns about are real ones. If the vector is highly structured, with a lot of zeros and a lot of repetitions of the same nonzero numbers, so it looks like it could be multiplied by a scalar and have small integer values, then it *probably* (cannot be certain) is a true direction of recession.

Directions of Recession and Constancy (cont.)

This test based on eigenvectors of the Fisher information matrix is not only inexact, even if some eigenvector is nearly along a direction of recession, this doesn't say which way is the direction of recession. (Directions of recession point one way. Eigenvectors don't. If v is an eigenvector, so is $-v$).

Directions of Recession and Constancy (cont.)

So suppose we have a submodel direction of recession δ_β.

Mapping to the saturated model, we get a direction of recession

$$\delta_\varphi = M\delta_\beta$$

We only care about the signs of components of δ_φ. If the j-th component of δ_φ is positive, then φ_j goes to $+\infty$ when the likelihood is maximized. And similarly for negative and $-\infty$. Only the zero components of δ_φ correspond to components of φ that stay finite.

Directions of Recession and Constancy (cont.)

Consider a single arrow, the j-th.

Suppose the one-parameter family for the arrow has convex support which is the interval from a_j to b_j (either of which can be infinite).

The inequalities this makes for the response vector of the aster model are

$$a_j y_{p(j)} \le y_j \le b_j y_{p(j)}$$

Since these involve at most two coordinates of the response vector, a direction of recession that yields an LCM that only conditions on $a_j y_{p(j)} = y_j$ or $y_j = b_j y_{p(j)}$ has at most two nonzero coordinates.

The direction of recession

$$\delta_{\varphi,k} = \begin{cases} -1, & k = j \\ a_j & k = p(j) \\ 0, & \text{otherwise} \end{cases}$$

yields the LCM that conditions on $a_j y_{p(j)} = y_j$.

Directions of Recession and Constancy (cont.)

The direction of recession

$$\delta_{\varphi,k} = \begin{cases} 1, & k = j \\ -b_j & k = p(j) \\ 0, & \text{otherwise} \end{cases}$$

yields the LCM that conditions on $y_j = b_j y_{p(j)}$.

More complicated directions of recession yield aster models with more arrows conditioned at their upper or lower bounds.

Directions of Recession and Constancy (cont.)

When we condition on one or more arrows being at one of their bounds, we have the same aster model we had before with the following changes.

- The j-th arrow now corresponds to the degenerate exponential family of distributions concentrated at a_j or b_j. We need to figure out its cumulant function.
- What was the direction of recession is now a direction of constancy. So we no longer have uniqueness of the MLE (in the limiting conditional model).

From now on we write b_j for either bound (lower or upper). Conditioning on the j-th arrow being at its bound we write as $y_j = b_j y_{p(j)}$ with b_j now standing for whichever bound we are conditioning on.

Degenerate One-Parameter Exponential Families

Suppose we have a one-parameter exponential family concentrated at the point b. What is its cumulant function?

The PMF is

$$f(y) = \begin{cases} 1, & y = b \\ 0, & \text{otherwise} \end{cases}$$

The only data we can observe is $y = b$, and for that the log likelihood is $\log(1) = 0$. And this does not depend on the parameter (all parameter values correspond to this same degenerate distribution). So

$$0 = l(\theta) = y\theta - c(\theta) = b\theta - c(\theta)$$

so we must have

$$c(\theta) = b\theta, \qquad \text{for all } \theta$$

Degenerate One-Parameter Exponential Families (cont.)

Let us check that the rest of the theory works too

$$\begin{aligned} c(\theta) &= b\theta \\ c'(\theta) &= b \\ c''(\theta) &= 0 \end{aligned}$$

which says the canonical statistic Y has mean b and variance 0, which is correct for the degenerate distribution concentrated at b.

Directions of Recession and Constancy (cont.)

Unfortunately, the R package `aster` does not allow degenerate distributions (concentrated at one point) for arrows.

The R package `aster2` does allow them, but is not ready for ordinary users.

So we need to figure out how a model with degenerate arrows corresponds to models without them.

Directions of Recession and Constancy (cont.)

$$\theta_j = \varphi_j + \sum_{\substack{k \in J \\ p(k) = j}} c_k(\theta_k) \qquad (*)$$

Recall (deck 2, slide 32) that $(*)$ must be used in an order that calculates θ_j for successors before θ_j for predecessors.

Suppose we are processing the j-th arrow, which is degenerate.

Directions of Recession and Constancy (cont.)

$$c_j(\theta_j) = b_j \theta_j$$

$$\begin{aligned}
\theta_j &= \varphi_j + \sum_{\substack{k \in J \\ p(k)=j}} c_k(\theta_k) \\
\theta_{p(j)} &= \varphi_{p(j)} + \sum_{\substack{m \in J \\ p(m)=p(j)}} c_m(\theta_m) \\
&= \varphi_{p(j)} + c_j(\theta_j) + \sum_{\substack{m \in J \\ p(m)=p(j) \\ m \neq j}} c_m(\theta_m) \\
&= \varphi_{p(j)} + b_j \left[\varphi_j + \sum_{\substack{k \in J \\ p(k)=j}} c_k(\theta_k) \right] + \sum_{\substack{m \in J \\ p(m)=p(j) \\ m \neq j}} c_m(\theta_m)
\end{aligned}$$

Directions of Recession and Constancy (cont.)

$$\theta_{p(j)} = \varphi_{p(j)} + b_j \varphi_j + b_j \sum_{\substack{k \in J \\ p(k) = j}} c_k(\theta_k) + \sum_{\substack{m \in J \\ p(m) = p(j) \\ m \neq j}} c_m(\theta_m)$$

For all of the distributions we have mentioned in the course b_j will be either zero or one.

- Bernoulli and Poisson have lower bound zero.
- Bernoulli has upper bound one.
- Zero-truncated Poisson has lower bound one.

If we are only dealing with these kinds of arrows then we always have $b_j = 0$ or $b_j = 1$ in the formula.

Directions of Recession and Constancy (cont.)

If $b_j = 0$, that is, we are conditioning on $Y_j = 0$, this essentially eliminates the j-th node and all of its successors, successors of successors, etc. from the model (we know they are all zero), and the formula on the preceding slide becomes

$$\theta_{p(j)} = \varphi_{p(j)} + \sum_{\substack{m \in J \\ p(m) = p(j) \\ m \neq j}} c_m(\theta_m)$$

just what we have when we set up the aster model with the j-th node and all of its successors, successors of successors, etc. eliminated.

Directions of Recession and Constancy (cont.)

If $b_j = 1$, that is, we are conditioning on $Y_j = Y_{p(j)}$, this essentially fuses Y_j and $Y_{p(j)}$ into one variable, and the formula from two slides ago becomes

$$\theta_{p(j)} = \varphi_{p(j)} + \varphi_j + \sum_{\substack{k \in J \\ p(k) = j \mathrm{or} p(k) = p(j)}} c_k(\theta_k)$$

This is just the formula we get if we fuse the j-th and $p(j)$-th nodes of the original model, hanging all of the successors of either j or $p(j)$ off of the fused node.

The canonical parameter for this fused node is $\varphi_{p(j)} + \varphi_j$ so the sum of the "regression equations" for each of the nodes that are fused applies to the fused node.

Directions of Recession and Constancy (cont.)

In either case ($b_j = 0$ or $b_j = 1$) this gives us a recipe for setting up an aster model which does have a maximum likelihood estimate and for which we can do inference.

But a bunch of issues remain. This tells us how to do inference for the LCM but we don't believe the MLE is the truth ($\hat{\beta}$ is not β).

So we don't believe the canonical parameter goes all the way to infinity and we don't believe the mean value parameter goes all the way to the boundary of the convex support.

Geyer (2009) describes how to do one-sided confidence intervals that address this issue, but the R package `aster` does not implement them and the R package `aster2` does not implement them yet.

Directions of Recession and Constancy (cont.)

The best we can do for now, and what everyone has done whenever this issue has arisen (whenever an actual direction of recession that was not a direction of constancy was discovered) is "fix up" the data by either deleting some nodes of the graph or fusing some nodes of the graph, thus forming the limiting conditional model (although users weren't always aware of that description of what they were doing).

Then we just analyze the "fixed up" data.

Example

Several real examples of directions of recession that are not directions of constancy have arisen in real data. But because the `aster` package does not handle them correctly, they have been treated as something of an embarrassment and only the “fixed up” data has been publicly analyzed.

The only published data that has directions of recession (as originally analyzed) is the aphid data that Shaw et al. (*American Naturalist*, 2008) to show how to do population growth rate reanalysis (and we redid but with a different submodel that doesn’t have directions of recession in Deck 4).

Example (cont.)

Rather than redo aphids, we will use some toy data.

```
> d<-"http://www.stat.umn.edu/geyer/8931aster/foobar.rda"
> load(url(d))
> rm(d)
> ls()

[1] "fam"    "pred"   "redata" "vars"
```

Example (cont.)

```
> vars

[1] "surv"        "has.flowers" "flowers"
[4] "seeds"

> pred

[1] 0 1 2 3

> fam

[1] 1 1 3 2

> sapply(redata, class)

      trt       blk      varb      resp        id
 "factor"  "factor"  "factor" "numeric" "integer"
     root       fit
"numeric" "numeric"
```

Example (cont.)

Everything is much the same as we expect for a long format aster dataset. The variables `varb`, `resp`, `id`, `root`, and `fit` are as usual with the latter being the indicator of “fitness” nodes, which are in this case the terminal nodes, the `"seeds"` ones.

The two categorical predictors

```
> levels(redata$trt)

[1] "a" "b" "c"

> levels(redata$blk)

[1] "A" "B" "C" "D"
```

Example (cont.)

```
> library(aster)
> aout <- aster(resp ~ varb + fit : (trt * blk), pred, fam,
+       id, root, data = redata)
> try(summary(aout))
```

```
apparent null eigenvectors of information matrix
directions of recession or constancy of log likelihood
 [1]  0.0000000  0.0000000  0.0000000  0.0000000
 [5]  0.3162278  0.0000000 -0.3162278 -0.3162278
 [9] -0.3162278  0.3162278  0.3162278  0.3162278
[13]  0.3162278  0.3162278  0.3162278
```

Oops! But in this example, we expect that!

Example (cont.)

```
> fred <- eigen(aout$fisher, symmetric = TRUE)
> dor <- fred$vectors[ , fred$values == min(fred$values)]
> names(dor) <- names(aout$coefficients)
> dor <- zapsmall(dor / max(dor))
> dor

    (Intercept) varbhas.flowers       varbseeds
              0               0               0
       varbsurv        fit:trta        fit:trtb
              0               1               0
       fit:blkB        fit:blkC        fit:blkD
             -1              -1              -1
   fit:trtb:blkB   fit:trtc:blkB   fit:trtb:blkC
              1               1               1
   fit:trtc:blkC   fit:trtb:blkD   fit:trtc:blkD
              1               1               1
```

Example (cont.)

Because there are so many nonzero components, this is very confusing.

But the fact that we can multiply the putative direction of recession by a scalar and get all the components to be small integers means this is almost certainly a true direction of recession.

```
> modmat <- aout$modmat
> dim(modmat)
```

```
[1] 300   4  15
```

```
> modmat <- as.vector(modmat)
> modmat <- matrix(modmat, ncol = length(dor))
> dor.phi <- modmat %*% dor
> dor.phi <- as.vector(dor.phi)
```

Example (cont.)

```
> unique(dor.phi)

[1] 0 1

> sum(dor.phi)

[1] 25

> foo <- data.frame(trt = as.character(redata$trt),
+     blk = as.character(redata$blk), id = redata$id,
+     varb = as.character(redata$varb),
+     resp = redata$resp, stringsAsFactors = FALSE)
> foo <- foo[dor.phi == 1, ]
```

Example (cont.)

```
> unique(foo$trt)

[1] "a"

> unique(foo$blk)

[1] "A"

> unique(foo$varb)

[1] "seeds"

> unique(foo$id)

 [1]   1  13  25  37  49  61  73  85  97 109 121 133
[13] 145 157 169 181 193 205 217 229 241 253 265 277
[25] 289
```

Example (cont.)

```
> unique(foo$resp)

[1] 0
```

So that's the story. Every individual in treatment "a" and block "A" had zero seeds.

What we do about it depends on what the scientific issues are.

If we took out the interaction, we wouldn't have a direction of recession.

But perhaps the interaction is the main issue of scientific interest.

Example (cont.)

If we collapsed some blocks, putting block `"A"` together with some other one, or if we collapsed some treatments, putting treatment `"a"` together with some other one, we wouldn't have a direction of recession.

But perhaps the changing the treatments or the blocks is also unacceptable scientifically.

Example (cont.)

We could just delete all individuals in treatment `"a"` and block `"A"` had zero seeds. They had zero observed fitness.

We just say that without doing any statistics about them.

We fit the aster model and do statistics about the rest.

This has the drawback that the deleted individuals do not contribute to the estimation of survival and number of flowers (which is, strictly speaking, wrong).

Example (cont.)

If all of these easy solutions to the problem are considered scientifically unacceptable, then the analysis becomes hard.

The R package `aster` insists that every individual have the same graph.

But we want individuals in treatment `"a"` and block `"A"` to have a different graph (with only three nodes not four, no `"seeds"`).

But the R package `aster` does not care what you call an individual. We can, if we like, treat the whole dataset as one individual.

This makes the graph a lot harder to specify.

Example (cont.)

```
> outies <- dor.phi == 1
> subdata <- redata[! outies, ]
```

We have now destroyed the structure of the aster model and must construct it anew.

```
> id <- subdata$id
```

This saves what the real individual numbers were.

```
> subdata$id <- 1
```

There is now just one individual. What is its graph?

Example (cont.)

```
> idx <- seq(1, nrow(subdata))
> varb <- as.character(subdata$varb)
> pred <- rep(NA, length(idx))
> fam <- rep(NA, length(idx))
> pred[varb == "surv"] <- 0
> fam[varb == "surv"] <- 1
> head(idx[varb == "surv"])

[1] 1 2 3 4 5 6

> head(idx[varb == "has.flowers"])

[1] 301 302 303 304 305 306
```

Example (cont.)

```
> sum(varb == "surv") == sum(varb == "has.flowers")

[1] TRUE

> pred[varb == "has.flowers"] <- idx[varb == "surv"]
> fam[varb == "has.flowers"] <- 1
> sum(varb == "has.flowers") == sum(varb == "flowers")

[1] TRUE

> pred[varb == "flowers"] <- idx[varb == "has.flowers"]
> fam[varb == "flowers"] <- 3
```

Example (cont.)

Now we get to the tricky bit (as if that wasn't tricky enough already).

```
> sum(varb == "flowers") == sum(varb == "seeds")

[1] FALSE

> bar <- match(id[varb == "seeds"], id[varb == "flowers"])
> pred[varb == "seeds"] <- idx[varb == "flowers"][bar]
> fam[varb == "seeds"] <- 2
```

Example (cont.)

Are we ready? No.

```
aout.sub <- aster(resp ~ varb + fit : (trt * blk),
    pred, fam, varb, id, root, data = subdata)
```

gives an error. It seems that the R function `aster` figures out the number of nodes from the unique elements of `varb`. So we have to make a correct `varb`.

```
> subvarb <- paste(as.character(subdata$varb), id,
+     sep = "")
> subdata <- data.frame(subdata, subvarb = subvarb)
```

Example (cont.)

Are we ready?

```
> aout.sub <- aster(resp ~ varb + fit : (trt * blk),
+     pred, fam, subvarb, id, root, data = subdata)
> summary(aout.sub)

Call:
aster.formula(formula = resp ~ varb + fit:(trt * blk), pred
    fam = fam, varvar = subvarb, idvar = id, root = root, d

                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.64088    0.05383  11.906  < 2e-16
varbhas.flowers -3.35703    0.25605 -13.111  < 2e-16
varbseeds       -0.02653    0.08143  -0.326   0.7446
varbsurv         0.23836    0.21093   1.130   0.2585
fit:trta        -0.17879    0.04409  -4.055 5.02e-05
fit:trtb        -0.04734    0.05481  -0.864   0.3877
fit:blkB         0.12805    0.07027   1.822   0.0684
fit:blkC         0.21500    0.06639   3.238   0.0012
```

Example (cont.)

```
> summary(aout.sub)


Call:
aster.formula(formula = resp ~ varb + fit:(trt * blk), pred = pred,
    fam = fam, varvar = subvarb, idvar = id, root = root, data = subdata)

                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.64088    0.05383  11.906  < 2e-16 ***
varbhas.flowers -3.35703    0.25605 -13.111  < 2e-16 ***
varbseeds       -0.02653    0.08143  -0.326   0.7446
varbsurv         0.23836    0.21093   1.130   0.2585
fit:trta        -0.17879    0.04409  -4.055 5.02e-05 ***
fit:trtb        -0.04734    0.05481  -0.864   0.3877
fit:blkB         0.12805    0.07027   1.822   0.0684 .
fit:blkC         0.21500    0.06639   3.238   0.0012 **
fit:blkD         0.20003    0.04569   4.378 1.20e-05 ***
fit:trtb:blkB   -0.05711    0.08816  -0.648   0.5171
fit:trtc:blkB   -0.02956    0.06696  -0.442   0.6588
fit:trtb:blkC   -0.10005    0.08363  -1.196   0.2316
fit:trtc:blkC   -0.11651    0.06252  -1.864   0.0624 .
fit:trtb:blkD   -0.01389    0.06693  -0.208   0.8356
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Original predictor variables dropped (aliased)
     fit:trtc
     fit:trtc:blkD
```

Example (cont.)

Check.

```
> mu <- predict(aout)
> mu.sub <- predict(aout.sub)
> all.equal(mu[outies], rep(0, sum(outies)))

[1] TRUE

> all.equal(mu[! outies], mu.sub)

[1] "Mean relative difference: 0.00462172"
```

Despite the not exact equality, it seems to be close enough to be a check. The `aster` function does not need to drive β all the way to infinity to claim convergence and quit iterating.

Summary

Was the example too simple or too complicated?

Too simple to show all the issues that arise.

More complicated than many users want to deal with or try to explain in a paper.

That is why we suggested 4 solutions to our toy problem. Sometimes changing the model or just eliminating some individuals from the data is the best way to go. Much easier to explain.

Summary (cont.)

So when you get the dreaded warning about directions of recession

- it may be a false positive that futzing with `info.tol` may reveal, or
- it may be a true positive that you have to actually deal with: identify the cause (what data is at what bound) and
  - change the submodel (one can always get rid of a direction of recession by fitting a simpler model with fewer parameters) or
  - change the data (there is always an LCM and one can always fit it by doing enough work)

  so that cause is eliminated.

Summary (cont.)

Either kind of solution, change the model or change the data, requires academic weasel wording in the write up.

Changing the model may be wrong because a simpler model that does not have a direction of recession

- does not fit the data as well (as shown by hypothesis tests) or
- does not address the issues of scientific interest.

Changing the data to the LCM is wrong because the LCM does not describe how close the canonical parameters of the original model are to infinity or how close the mean value parameters of the original model are to the boundary of the convex support.

In short, analysis of the LCM tells you anything statistics can tell you about the LCM. What it doesn't tell you is how close the true unknown mean values of the data the LCM fixes at the boundary are to really being at the boundary.

Other Issues

If you do a likelihood ratio test (with `anova.asterOrReaster`) and the smaller model has no directions of recession, the test is valid (regardless of whether the larger model has directions of recession). Geyer (2009, Section 3.15) explains.

If you do a likelihood ratio test (with `anova.asterOrReaster`) and the smaller model has directions of recession, the test is invalid. The likelihood ratio test statistic is approximately chi-squared but the degrees of freedom needs to be calculated differently.

If you do a likelihood ratio test (with `anova.asterOrReaster`) applied the the LCM (constructed as we did in the example) so the null hypothesis applied to the LCM data has no directions of recession, the test is valid.

Other Issues (cont.)

Confidence intervals (Geyer, 2009, Section 3.16) are even more complicated.

The only principle that is simple to understand is (repeating what was said earlier) statistical analysis of the LCM

- does give valid inference about the parameters of the LCM,
- does not give valid inference (or any inference) about the parameters of the original model that are gone in the LCM.

What are those parameters that are "gone"? Easiest to see for conditional mean value parameters: those for the arrows that have been removed or fused. Hard to see for canonical parameters because they are all mixed up. Some directions in the canonical parameter space (the directions of constancy of the LCM) are "gone" in the LCM.

One-Sided Confidence Intervals

Here is a simple idea from Geyer (2009) that is the basis of all the one-sided confidence intervals proposed therein.

Suppose we have binomial data and we want to test

$$
\begin{aligned}
H_0 &: p = p_0 \\
H_1 &: p < p_0
\end{aligned}
$$

that is, a simple lower-tailed test.

The obvious P-value is

$$\operatorname{pr}_{p_0}(X \le x),$$

where x is the observed value of the binomial data and X is a random variable having the null distribution of the test statistic, which here is $\text{Binomial}(n, p_0)$.

So far, standard elementary statistics.

One-Sided Confidence Intervals (cont.)

Now we want to invert the level α one-tailed hypothesis test to make a one-sided $1 - \alpha$ confidence interval.

The interval is all of the p_0 that the test does not reject at level α, and the test rejects $H_0 : p = p_0$ at level α when $P \leq \alpha$, that is, when

$$\text{pr}_{p_0}(X \leq x) \leq \alpha$$

so the corresponding confidence interval is

$$\{\, p \in [0, 1] : \text{pr}_p(X \leq x) \geq \alpha \,\}$$

One-Sided Confidence Intervals (cont.)

Now specialize to the case where we observe $x = 0$. The one-sided $1 - \alpha$ confidence interval is

$$\{\, p \in [0, 1] : \operatorname{pr}_p(X = 0) \geq \alpha \,\}$$

that is, we need p such that

$$(1 - p)^n \geq \alpha$$

and that interval is

$$0 \leq p \leq 1 - \alpha^{1/n}$$

One-Sided Confidence Intervals (cont.)

Similar logic works for any discrete distribution. For Poisson, when we observe $x = 0$, the interval is

$$\{\, \mu \in [0, \infty) : \operatorname{pr}_\mu(X = 0) \geq \alpha \,\}$$

that is, we need μ such that

$$e^{-n\mu} \geq \alpha$$

and that interval is

$$0 \leq \mu \leq -\frac{\log(\alpha)}{n}$$

(this is for observing n IID Poisson(μ) individuals).

Example (cont.)

In our example we had

```
> idout <- redata$id[outies]
> rowout <- redata$id %in% idout
> varbflowers <- as.character(redata$varb) == "flowers"
> nzero <- sum(redata$resp[rowout & varbflowers])
> nzero
```

```
[1] 38
```

flowers observed in the class (treatment `"a"` and block `"A"`) in which zero seeds were observed.

Thus we have predecessor `nzero` and successor zero for a Poisson arrow.

Example (cont.)

We want to make a one-sided interval for the conditional mean (ξ_j not μ_j) number of seeds in this class in which zero seeds were observed. Thus `nzero` is the *n* for this procedure.

We assumed seed count was (conditionally) Poisson. Thus the corresponding one-sided confidence interval is

```
> conf.level <- 0.95
> alpha <- 1 - conf.level
> c(0, - log(alpha) / nzero)
```

```
[1] 0.00000000 0.07883506
```

One-Sided Confidence Intervals (cont.)

So this is in one sense the usual story. In this class we have $\hat{\xi} = 0$ but we don't make the elementary mistake of confusing the sample and the population, of confusing $\hat{\xi}$ and ξ.

Our one-sided 95% confidence interval (0, 0.08) is not taught in intro stats, but is not rocket science.

One-Sided Confidence Intervals (cont.)

But any further analysis becomes very complicated very fast, and we have not thought of any way to make it simple (there may be no way to make it simple).

Example (cont.)

```
> iout <- redata$id[outies]
> mu.sub.too <- predict(aout.sub, se.fit = TRUE)
> fred <- id %in% iout & subdata$varb %in% "flowers"
> mu.hat <- unique(mu.sub.too$fit[fred])
> se.mu.hat <- unique(mu.sub.too$se.fit[fred])
> mu.hat

[1] 0.2379846

> se.mu.hat

[1] 0.04222433
```

Example (cont.)

So that gives us a 95% asymptotic confidence interval for flower count in this class

```
> zcrit <- qnorm((1 + conf.level) / 2)
> mu.hat + c(-1, 1) * zcrit * se.mu.hat
```

```
[1] 0.1552265 0.3207428
```

We know

$$\mu_j = \xi_j \mu_{p(j)}$$

(deck 2, slide 74) and now we have confidence intervals for ξ_j and $\mu_{p(j)}$. Can we put them together?

Example (cont.)

With a little thought it becomes clear that we want to combine two one-sided intervals (no point in combining a one-sided and a two-sided).

```
> zcrit <- qnorm(conf.level)
> u1 <- - log(alpha) / nzero
> u2 <- mu.hat + zcrit * se.mu.hat
> c(0, u1)

[1] 0.00000000 0.07883506

> c(0, u2)

[1] 0.0000000 0.3074375

> c(0, u1 * u2)

[1] 0.00000000 0.02423685
```

One-Sided Confidence Intervals (cont.)

Since the intervals we combined did not have simultaneous coverage, we only get a 90% confidence interval (this is Bonferroni correction: add the alphas not the confidence levels).

Summary (cont.)

All of this can become arbitrarily complicated in an aster model with a complicated graph and several arrows being conditioned on in the LCM.

We do not have functions to deal with this, mainly because it is not clear what users will want in complicated situations.

Honesty compells me to add that I do not know what happens in all complicated situations. Geyer (2009) has a complete analysis of what can happen in GLM and log-linear models for categorical data. No such complete analysis has been done for aster models (for all possible canonical affine submodels, what are all possible LCM). So each new rigorous analysis may bring surprises. I didn't know how the example done in this deck of slides would work until I worked through it.